SIAM J. APPL. MATH.
Vol. 76, No. 1, pp. 197–216

SOJOURN TIME IN A SINGLE-SERVER QUEUE WITH THRESHOLD SERVICE RATE CONTROL*

IVO ADAN† AND BERNARDO D'AURIA‡

Abstract. We study the sojourn time in a queueing system with a single exponential server, serving a Poisson stream of customers in order of arrival. Service is provided at a low or high rate, which can be adapted at exponential inspection times. When the number of customers in the system is above a given threshold, the service rate is upgraded to the high rate, otherwise, it is downgraded to the low rate. The state dependent changes in the service rate make the analysis of the sojourn time a challenging problem, since the sojourn time now also depends on future arrivals. We determine the Laplace transform of the stationary sojourn time and describe a procedure to compute all moments as well. First we analyze the special case of continuous inspection, where the service rate immediately changes once the threshold is crossed. Then we extend the analysis to random inspection times. This extension requires the development of a new methodological tool, that is, *matrix generating functions*. The power of this tool is that it can also be used to analyze generalizations to phase-type services and inspection times.

Key words. sojourn time distribution, matrix generating function, adaptable service speed

AMS subject classifications. 60K25, 60K37, 60D05

DOI. 10.1137/14097046X

1. Introduction.

We consider a single-server queueing system, where customers arrive according to a Poisson stream with rate λ and receive service in order of arrival. The service requirements are exponential with mean 1. The rate of the server can be either μ_0 or μ_1 and this service rate can be adapted at random inspection times that occur according to a Poisson stream with rate γ. For convenience, we think of μ_1 as the fastest rate, i.e., $\mu_1 > \mu_0$, even if under the stability conditions this assumption may be removed. When the number of customers in the system is above the threshold K, the service rate is upgraded to the high rate μ_1, otherwise, it is downgraded to the low rate μ_0. An important performance measure is the sojourn time. In this paper we aim to determine its stationary distribution. This is a challenging problem, since due to adaptable service rate, the sojourn time does not only depend on the state seen at arrival, but it also depends on future arrivals.

There is a considerable literature on the analysis of single-server queueing systems with variable service rates; see, e.g., [1, 2, 6, 7, 10, 11]. Those studies often assume that the service rates can be continuously adapted based on the queue content, and focus on the calculation of the steady-state workload distribution. An exponential multi-server system is considered in [14], with the feature that a reserved block of servers can be switched on (which takes an exponential switch-on time) when the number of customers in the system exceeds a certain threshold, and this block is immediately

*Received by the editors May 27, 2014; accepted for publication (in revised form) September 29, 2015; published electronically January 28, 2016.
http://www.siam.org/journals/siap/76-1/97046.html
†Mechanical Engineering Department, Technische Universiteit Eindhoven, Postbus 513, 5600 MB Eindhoven (iadan@tue.nl).
‡Statistics Department, Madrid University Carlos III, Avda. Universidad, 30. 28911 Leganés (Madrid) Spain (bernardo.dauria@uc3m.es). This author's research was partially supported by the Spanish Ministry of Education and Science Grants MTM2010-16519, SEJ2007-64500, MTM2013-42104-P (FEDER funds); and the Dutch Star grant of October 2013. The second author wants to thank the research institutes ICMAT (Madrid, Spain) and EURANDOM (Eindhoven, The Netherlands) for kindly hosting him during the development of this project.

switched off when the number drops below another threshold. The emphasis in [14] is on the trade-off between the mean sojourn time and operating costs of the servers. In [3], the stationary distribution of the workload is determined for an $M/G/1$ queue, where the service rate cannot be continuously adapted, but only right after customer arrivals. In the literature, systems with adaptable service speed at inspection times have already been analyzed; we refer the reader to [4, 5] and the references therein.

The model with inspection rate $\gamma < \infty$ can be handled by considereing a two-stage birth-death process. This kind of model usually shows up in the analysis of retrial queues, where the state of the system has to keep track of the size of the retrial orbit. We refer the reader to the survey [9]. In [8], the number of retrials of a generic customer is analyzed, which is a quantity directly related to the sojourn time and which depends on future arrivals to the system. Falin [8] starts the analysis with a matrix equation that is similar to the one appearing in section 3, but is able to reduce this equation to a scalar one by exploiting the fact that the retrial system has no buffer and only one server. Multichannel systems are much more complicated to analyze and very few results are available about the sojourn time. Generally, what makes retrial systems more complicated than the one analyzed here, is the property that the rate at which retrial customers arrive at the system is proportional to the size of the orbit. This phenomenon does not appear in our system, which is one of the reasons why our analysis is feasible.

As mentioned above, the focus in the current paper is on the sojourn time, not on the workload or number of customers in the system. In section 2, we first consider the special case of continuous inspection (so $\gamma = \infty$), where the service rate immediately changes once the threshold K is crossed. This assumption simplifies the model, though it still contains the complication that the sojourn time depends on future arrivals. For the case of continuous inspection, we determine the Laplace transform of the stationary sojourn time and describe a procedure to compute all moments as well. The computation of the Laplace transform requires a recursive scheme and for the case $\gamma < \infty$ the Laplace transform can be expressed in terms of matrix functions that can be computed as solutions of a linear matrix system.

Then, in section 3, we proceed by extending the analysis to random inspection times occurring according a Poisson stream with rate $\gamma < \infty$. This extension, however, is not straightforward, and it requires the development of a new methodological tool, that is, *matrix generating functions*. By employing this tool we are able to find an expression for the Laplace transform of the stationary sojourn time, involving finitely many terms which can be recursively calculated. The analytical results are illustrated by numerical examples.

2. Model with continuous inspection. In this section we first consider the special case of continuous inspection, so $\gamma = \infty$. This implies that whenever the number of customers in the system exceeds the threshold $K > 0$, the rate of the server is immediately upgraded from the low rate μ_0 to the high rate $\mu_1 > \mu_0$. As soon as the number of customers in the system becomes less than or equal to K, the rate of the server is reduced to the low rate μ_0 again.

Denoting by $Q(t)$, the number of customers in the system at time $t > 0$, we have that the process is a continuous time Markov chain, the transition diagram of which is depicted in Figure 1.

Denoting by Q^* the stationary number of customers in the system, we have that its distribution is given by

$$\pi_n = \mathbb{P}(Q^* = n) = \begin{cases} (\lambda/\mu_0)^n \, \pi_0 & \text{for } n \le K\,, \\ (\mu_1/\mu_0)^K \, (\lambda/\mu_1)^n \, \pi_0 & \text{for } n > K\,, \end{cases} \tag{2.1}$$

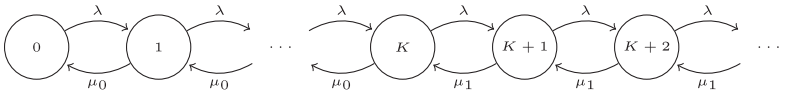

FIG. 1. *Transition diagram for continuous inspection model.*

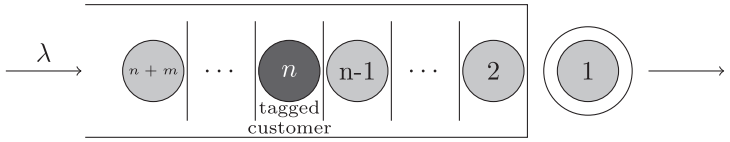

FIG. 2. *Tagged customer* (n, m) *at position* n *in the queue, with* m *customers behind him.*

and under the stability assumption $\mu_1 > \lambda$, the value of π_0 is given by

$$\pi_0 = \left(\sum_{n=0}^{K} (\lambda/\mu_0)^n + \frac{\lambda}{\mu_1 - \lambda} (\lambda/\mu_0)^K \right)^{-1} . \tag{2.2}$$

We aim to compute the distribution of the sojourn time of a typical customer that arrives at the system in a stationary regime. Note that, in order to do this, we cannot use Little's distributional law [12], since future arrivals may affect the sojourn times of the customers already present in the system by inducing a change in the service rate.

As shown in Figure 2, we identify a *tagged* customer in the queue by a pair of numbers (n, m), where n stands for the position of the tagged customer in the queue, and where m denotes the number of customers behind him. We denote the sojourn time of this customer (n, m) by $S(n, m)$. The stationary sojourn time is denoted by S^*.

For the Laplace transforms $\psi(s) = \mathbb{E}[e^{-s\,S^*}]$ and $\psi(n, m, s) = \mathbb{E}[e^{-s\,S(n,m)}]$, the following relation holds by virtue of PASTA [16],

$$\psi(s) = \sum_{n=0}^{\infty} \psi(n+1, 0, s)\, \pi_n \,. \tag{2.3}$$

Hence, to compute the Laplace transform of the stationary sojourn time S^*, we need to compute the transforms $\psi(n, 0, s)$ for each $n \geq 0$.

By using next-event analysis we have, for $n > 0$ (w.p. stands for with probability),

$$S(n,m) = \begin{cases} \frac{1}{\lambda+\mu_0} X + \begin{cases} S(n-1,m) & \text{w.p. } \mu_0/(\mu_0+\lambda), \\ S(n,m+1) & \text{w.p. } \lambda/(\mu_0+\lambda), \end{cases} & \text{as } n+m \leq K \,, \\ \frac{1}{\lambda+\mu_1} X + \begin{cases} S(n-1,m) & \text{w.p. } \mu_1/(\mu_1+\lambda), \\ S(n,m+1) & \text{w.p. } \lambda/(\mu_1+\lambda), \end{cases} & \text{as } n+m > K \,, \end{cases} \tag{2.4}$$

where X denotes an independent exponential random variable with rate 1, and $S(0, m) = 0$. By Laplace transforming the relations (2.4), we get, for $n > 0$,

$$\psi(n, m, s) = \frac{\mu_{1\{n+m>K\}}\, \psi(n-1, m, s) + \lambda\, \psi(n, m+1, s)}{\lambda + \mu_{1\{n+m>K\}} + s} \tag{2.5}$$

with boundary conditions, $\psi(0, m, s) = 1$, for all $m \geq 0$, and where we used the indicator function $1\{A\} = 1$ if A is true and 0 otherwise.

When $m \geq K$, it follows that for any $n > 0$, the server will work at high speed during the whole sojourn time of the (n, m)-tagged customer. Hence $S(n, m)$ is Erlang distributed with parameters n and μ_1, and thus its Laplace transform is equal to

$$\psi(n, m, s) = (\mu_1/(\mu_1 + s))^n \quad \text{as } n > 0 \text{ and } m \geq K. \tag{2.6}$$

The above equation is also valid for $n = 0$.

Using expression (2.6) in (2.5), the Laplace transforms $\psi(n, m, s)$ for $m < K$ can be recursively computed in n, as the following lemma shows. The proof of the lemma is deferred to Appendix A.

LEMMA 2.1. *By defining*

$$\begin{aligned} a_s(k) &= \mu_0/(\lambda + \mu_0 + s)1\{k \leq K\} + \mu_1/(\lambda + \mu_1 + s)1\{k > K\}\,, \\ b_s(k) &= \lambda/(\lambda + \mu_0 + s)1\{k \leq K\} + \lambda/(\lambda + \mu_1 + s)1\{k > K\}\,, \end{aligned}$$

and $B_s(k, 0) = 1$ *and* $B_s(k, h + 1) = B_s(k, h)\, b_s(k + h)$ *for* $k, h \geq 0$*, we have*

$$\begin{aligned} \psi(n, m, s) &= B_s(n + m, K - m) \left(\frac{\mu_1}{\mu_1 + s}\right)^n \\ &\quad + \sum_{k=m}^{K-1} a_s(n + k)\, B_s(n + m, k - m)\, \psi(n - 1, k, s)\,, \end{aligned} \tag{2.7}$$

for $n > 0$ *and* $0 \leq m < K$*.*

Remark 2.2. It can be easily shown that the value of $B_s(k, h)$ can be explicitly computed by the following formula

$$B_s(k, h) = \left(\frac{\lambda}{s + \lambda + \mu_1}\right)^h \left(\frac{s + \lambda + \mu_1}{s + \lambda + \mu_0}\right)^{h \wedge (K-k+1)^+} \tag{2.8}$$

with $a \wedge b = \min\{a, b\}$ and $(a)^+ = \max\{a, 0\}$.

Relation (2.6) and Lemma 2.1 allow us to compute $\psi(n, m, s)$ for any $m, n \geq 0$. However, to calculate $\psi(s)$ in (2.3) we still need to compute an infinite number of terms. To overcome this issue we take advantage of the fact that, above the threshold K, the transition diagram is invariant towards the right, similarly to the standard $M/M/1$ queue. To use this invariant property we introduce the following marginal z-transform

$$\phi(z, m, s) = \sum_{h=0}^{\infty} \psi(K + h + 1, m, s)\, z^h\,, \tag{2.9}$$

valid for $|z| < 1$. In the following we show how to compute, in finitely many steps, the function $\phi(z, m, s)$. We use it to calculate the infinite sum in (2.3) and then obtain a formula to compute the Laplace transform of the sojourn time as given in Proposition 2.3.

By writing (2.5) for $n = K + h + 1$, multiplying by z^h, and summing over all $h \geq 0$, the following recursive equation holds:

$$\phi(z, m, s) = \frac{\mu_1\, \psi(K, m, s)}{\lambda + \mu_1(1 - z) + s} + \frac{\lambda\, \phi(z, m + 1, s)}{\lambda + \mu_1(1 - z) + s}\,. \tag{2.10}$$

As boundary value we have

$$\phi(z,K,s) = \sum_{h=0}^{\infty} \left(\frac{\mu_1}{\mu_1+s}\right)^{K+h+1} z^h = \left(\frac{\mu_1}{\mu_1+s}\right)^{K+1} \sum_{h=0}^{\infty} \left(\frac{\mu_1\, z}{\mu_1+s}\right)^h$$
$$= \left(\frac{\mu_1}{\mu_1+s}\right)^{K+1} \frac{\mu_1+s}{\mu_1(1-z)+s}\,, \tag{2.11}$$

from which the values of $\phi(z,m,s)$ can be recursively computed for $m = K-1,\dots,0$, yielding

$$\phi(z,m,s) = \sum_{h=0}^{K-1-m} \frac{\mu_1\,\lambda^h}{(\lambda+\mu_1(1-z)+s)^{h+1}} \psi(K,m+h,s)$$
$$+ \left(\frac{\lambda}{\lambda+\mu_1(1-z)+s}\right)^{K-m} \phi(z,K,s)\,. \tag{2.12}$$

In particular we can compute, in finitely many steps, the value of $\phi(z,0,s)$.

Knowing the value of $\phi(z,0,s)$, expression (2.3) can be finally computed as summarized in the following proposition.

PROPOSITION 2.3. *The Laplace transform of S^* can be computed in the form*

$$\psi(s) = \pi_0 \sum_{h=0}^{K-1} \left[\left(\frac{\lambda}{\mu_0}\right)^h \psi(h+1,0,s) + \left(\frac{\lambda}{\mu_0}\right)^K \left(\frac{\lambda}{\mu_1}\right)^h \left(\frac{\mu_1}{\mu_1+s}\right)^{h+1} \psi(K,h,s)\right]$$
$$+ \pi_0 \left(\frac{\lambda}{\mu_0}\right)^K \left(\frac{\lambda}{\mu_1}\right)^K \frac{\mu_1}{\mu_1-\lambda} \left(\frac{\mu_1}{\mu_1+s}\right)^{2K} \frac{\mu_1-\lambda}{\mu_1-\lambda+s}\,. \tag{2.13}$$

Proof. The result follows from (2.3) by splitting the sum into a finite part, $n < K$, and an infinite part,

$$\psi(s) = \pi_0 \sum_{n=0}^{K-1} \left(\frac{\lambda}{\mu_0}\right)^n \psi(n+1,0,s) + \pi_0 \left(\frac{\lambda}{\mu_0}\right)^K \phi(\lambda/\mu_1,0,s)\,. \tag{2.14}$$

For the last term we use (2.12) and (2.11) to get

$$\phi(\lambda/\mu_1,m,s) = \sum_{h=0}^{K-1-m} \left(\frac{\lambda}{\mu_1}\right)^h \left(\frac{\mu_1}{\mu_1+s}\right)^{h+1} \psi(K,m+h,s)$$
$$+ \frac{\mu_1}{\mu_1-\lambda+s} \left(\frac{\lambda}{\mu_1}\right)^{K-m} \left(\frac{\mu_1}{\mu_1+s}\right)^{2K-m} \tag{2.15}$$

and the result follows by rearranging terms. □

The terms appearing in (2.13) have the following nice probabilistic interpretation.

- With probability $\pi_h = \pi_0(\lambda/\mu_0)^h$, $h < K$, the tagged user enters a system with h customers and experiences a sojourn time, the Laplace transform of which is $\psi(h+1,0,s)$.
- With probability $\pi_{K+h} = \pi_0(\lambda/\mu_0)^K(\lambda/\mu_1)^h$, $0 < h < K$, he finds $K+h$ customers waiting. We slightly modify the system and assume that the tagged customer overtakes $h+1$ customers and occupies position K instead of the last

one in the queue. In addition, the server first serves the last $h+1$ customers. Since the speed of the server depends on the number of customers waiting and not on their specific order of service, the first $h+1$ services will be at speed μ_1 taking an Erlang time with parameters $h+1$ and μ_1 to complete. What is left is the service time of the tagged customer, the Laplace transform of which is $\psi(K,h,s)$.

- With probability $\pi_{\geq 2K} = \pi_0(\lambda/\mu_0)^K(\lambda/\mu_1)^K(\mu_1/(\mu_1-\lambda))$ the tagged customer finds at least $2K$ customers waiting. As before, he is going to occupy position K, the sojourn time of which is Erlang distributed with parameters K and μ_1. The number of customers he has overtaken is at least K, and the time it takes to complete their services is the sum of K exponential random variables with parameter μ_1 plus a generic sojourn time of an $M/M/1$ queue having μ_1 as service speed. This last quantity is exponentially distributed with parameter $\mu_1-\lambda$.

Remark 2.4. From the Laplace transform of the stationary sojourn time given in (2.3) an explicit expression for the distribution can be obtained. Indeed, the inverse transformation is straightforward as the Laplace transform is a rational polynomial, the poles of which are all located on the real axis. To be more precise, the locations of the poles belong to the set

$$\mathcal{A} = \{-(\lambda+\mu_1),\, -(\lambda+\mu_0),\, -\mu_1,\, -(\mu_1-\lambda)\}\,,$$

implying that the density function is given by a linear combination of terms $t^k e^{a\,t}$ for $a \in \mathcal{A}$ and $k = 0, 1, \ldots, \text{mult}(a)-1$, where $\text{mult}(a)$ denotes the multiplicity of pole a.

Remark 2.5. If we let $K \to \infty$ in (2.13), we recover (2.3). For any $n \geq 0$, $\psi(n+1,0,s)$ becomes the Laplace transform of an Erlang distribution with parameters $n+1$ and μ_0, and $\psi(s)$ reduces to the Laplace transform of an exponential distribution with parameter $\mu_0-\lambda$, that is the distribution of the sojourn time of a classical $M/M/1$ queue with service rate μ_0.

Remark 2.6. If $K=0$, only the last term in (2.13) is different from zero. Substituting $\pi_0 = (\mu_1-\lambda)/\mu_1$, given in (2.2), we get that $\psi(s)$ is the Laplace transform of an exponential distribution with parameter $\mu_1-\lambda$, that is, the distribution of the sojourn time of a classical $M/M/1$ queue with service rate μ_1.

2.1. First moment calculation. As mentioned in Remark 2.4, it is possible to compute the distribution of the sojourn time, but it is easier to compute the moments by using the relation $\mathbb{E}[S^k] = (-1)^k \psi^{(k)}(0+)$. In this section we show how to compute the first moment. However, by taking higher order derivatives of the Laplace transform, recursive expressions can be obtained to compute all moments.

Let $\nu = \mathbb{E}[S]$ and $\nu_{n,m} = \mathbb{E}[S(n,m)]$. With $n>0$, from (2.6) we have for $m \geq K$, $\nu_{n,m} = n/\mu_1$, and using (2.7), for $0 \leq m < K$,

$$\begin{aligned}\nu_{n,m} = {} & \frac{n}{\mu_1} B_{0^+}(n+m, K-m) - B'_{0^+}(n+m, K-m) \\ & + \sum_{k=m}^{K-1} \Big(\nu_{n-1,k}\, a_{0^+}(n+k)\, B_{0^+}(n+m, k-m) \\ & \qquad - a'_{0^+}(n+k)\, B_{0^+}(n+m, k-m) \\ & \qquad - a_{0^+}(n+k)\, B'_{0^+}(n+m, k-m) \Big)\,, \end{aligned} \tag{2.16}$$

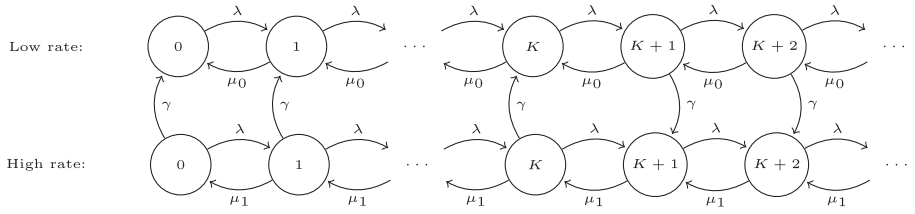

FIG. 3. *Transition diagram for exponential inspection times.*

where

$$a'_{0^+}(k) = -\mu_0/(\lambda+\mu_0)^2 1\{k \le K\} - \mu_1/(\lambda+\mu_1)^2 1\{k > K\}\,,$$
$$b'_{0^+}(k) = -\lambda/(\lambda+\mu_0)^2 1\{k \le K\} - \lambda/(\lambda+\mu_1)^2 1\{k > K\}\,,$$

and $B'_{0^+}(k,0) = 0$ and $B'_{0^+}(k,h+1) = B'_{0^+}(k,h)\, b_0(k+h) + B_0(k,h)\, b'_{0^+}(k+h)$.

The following algorithm shows how to recursively compute $\nu_{n,m}$ for $0 \le m < K$:

ALGORITHM 1. Computing $\nu_{n,m}$ for $n > 0$ and $0 \le m < K$

for i=1 **to** n **do**
  for j=1 **to** K-m **do**
    compute $\nu_{i,K-j}$
  end
end

Finally, by applying Proposition 2.3, we get

$$\nu = \pi_0 \sum_{h=0}^{K-1} \left[\left(\frac{\lambda}{\mu_0}\right)^h \nu_{h+1,0} + \frac{\lambda^K}{\mu_0^K} \left(\frac{\lambda}{\mu_1}\right)^h \left(\nu_{K,h} + \frac{h+1}{\mu_1}\right) \right]$$
$$+ \pi_0 \left(\frac{\lambda}{\mu_0}\right)^K \left(\frac{\lambda}{\mu_1}\right)^K \left(\frac{2K}{(\mu_1-\lambda)} + \frac{\mu_1}{(\lambda-\mu_1)^2}\right) . \tag{2.17}$$

3. Model with inspection times. In this section we analyze the system where inspection times occur according to a Poisson stream with rate $\gamma < \infty$. So, in this case, there is no continuous inspection and adaptation of the service rate is delayed (with an exponential time) when the number of customers in the system crosses the threshold K. If at an inspection time the system is found congested with more than K customers, the service rate is immediately set to the fast rate μ_1 and otherwise, if at most K customers are present, the service rate is set to the low rate μ_0.

Now we need to include the service rate in the state description of the system, resulting in the Markov chain shown in Figure 3. Note that for any number of customers in the system, the service rate can be high and low.

Denoting by $\mathcal{M}$ the stationary random service rate, let $\pi_{0n} = \mathbb{P}(\mathcal{M} = \mu_0, Q^* = n)$ and $\pi_{1n} = \mathbb{P}(\mathcal{M} = \mu_1, Q^* = n)$ be the stationary probabilities to find n customers in the system with the server working at rates μ_0 and μ_1, respectively. In what follows, the quantity π_n denotes the column vector with components $(\pi_{0n}, \pi_{1n})^\top$, where $(\cdot)^\top$ is the transposition operator. The stationary distribution satisfies the balance equations

$$\begin{aligned} -H_1\,\pi_0 + M\,\pi_1 &= 0, \\ \Lambda\,\pi_{n-1} - H_2\,\pi_n + M\,\pi_{n+1} &= 0, \qquad 1 \le n \le K\,, \\ \Lambda\,\pi_{n-1} - H_3\,\pi_n + M\,\pi_{n+1} &= 0, \qquad n > K\,, \end{aligned} \tag{3.1}$$

where the transition matrices are defined by

$$M = \begin{pmatrix} \mu_0 & 0 \\ 0 & \mu_1 \end{pmatrix} , \quad \Lambda = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} ,$$

and $H_1 = \Lambda + \Gamma_2$, $H_2 = M + \Lambda + \Gamma_2$, and $H_3 = M + \Lambda + \Gamma_3$, where

$$\Gamma_2 = \begin{pmatrix} 0 & -\gamma \\ 0 & \gamma \end{pmatrix} , \quad \Gamma_3 = \begin{pmatrix} \gamma & 0 \\ -\gamma & 0 \end{pmatrix} .$$

From the theory on quasi-birth-death processes [15, 13], we conclude that for $n > K$, the stationary probability vector π_n can be written in the form

$$\pi_{K+h} = R^h \, \pi_K, \quad h \geq 0 , \tag{3.2}$$

where the matrix R is the minimal nonnegative solution of the matrix equation

$$\Lambda - H_3 \, R + M \, R^2 = 0 . \tag{3.3}$$

Using the probabilistic interpretation of R, or by solving the matrix equation (3.3), it follows that R is of triangular form and, in particular, it is equal to

$$R = \begin{pmatrix} R_{00} & 0 \\ \frac{\gamma}{\mu_1}\frac{R_{00}}{1-R_{00}} & \frac{\lambda}{\mu_1} \end{pmatrix} \tag{3.4}$$

with $R_{00} = \frac{\mu_0+\gamma+\lambda}{2\mu_0} - \sqrt{(\frac{\mu_0+\gamma+\lambda}{2\mu_0})^2 - \frac{\lambda}{\mu_0}}$.

The value of π_K can be computed by the normalizing equation

$$\sum_{k=0}^{K-1} e \, \pi_k + e \, (I - R)^{-1} \pi_K = 1 \tag{3.5}$$

with e the all-one (row) vector.

By PASTA, as in (2.3), the Laplace transform of the stationary sojourn time is given by

$$\psi(s) = \sum_{n=0}^{\infty} \psi(n+1, 0, s) \, \pi_n , \tag{3.6}$$

where $\psi(n, m, s)$ denotes the row vector $(\psi_0(n, m, s), \psi_1(n, m, s))$ with $\psi_i(n, m, s)$ being the Laplace transform of the sojourn time $S_i(n, m)$ of a tagged customer who is at position (n, m) and the service rate is μ_i, $i = 0, 1$.

By using next-event analysis, we get the following recursive equations for the sojourn times, $S_i(n, m)$, $i = 0, 1$, $n > 0$:

$$S_i(n,m) = \frac{X}{\lambda + \mu_i + \gamma} + \begin{cases} S_i(n-1, m) & \text{w.p. } \mu_i/(\lambda + \mu_i + \gamma) , \\ S_i(n, m+1) & \text{w.p. } \lambda/(\lambda + \mu_i + \gamma) , \\ S_{1\{n+m>K\}}(n, m) & \text{w.p. } \gamma/(\lambda + \mu_i + \gamma) , \end{cases} \tag{3.7}$$

where X denotes an independent exponential random variable with rate 1, and $S_i(0, m)$ $= 0$. Taking the Laplace transform of (3.7) yields the equation

$$\psi(n, m, s) \, (H(s) - \Gamma_{1\{n+m>K\}}) = \psi(n-1, m, s) \, M + \psi(n, m+1, s) \, \Lambda \tag{3.8}$$

for $n > 0$, where

$$H(s) = (\gamma + s) \, I + \Lambda + M , \quad \Gamma_0 = \begin{pmatrix} \gamma & \gamma \\ 0 & 0 \end{pmatrix} , \quad \Gamma_1 = \begin{pmatrix} 0 & 0 \\ \gamma & \gamma \end{pmatrix} ,$$

and $\psi(0, m, s) = e$, with e the all-one (row) vector and I the identity matrix.

Similarly to (2.6), when $m \geq K$ and for any $n > 0$, we have that whenever an inspection occurs, the service rate is set and kept at the value μ_1 till the end of the sojourn time of the tagged customer. This implies that $\psi(n, m+1, s) = \psi(n, m, s)$ for $m \geq K$ and $n > 0$, and substitution in (3.8) yields

$$\psi(n, m, s) = e\, T^n(s) \quad \text{as } n > 0 \text{ and } m \geq K \ , \tag{3.9}$$

with $T(s) = M\,(H(s) - \Gamma_1 - \Lambda)^{-1}$.

Equations (3.9) and (3.8) allow us to get the values of $\psi(n, m, s)$ for any $n, m \geq 0$. However to compute expression (3.6) in finitely many steps, we still need to find a way to handle the infinite sum. So far, the analysis proceeds as in section 2, and thus the next step would be to introduce the marginal z-transforms corresponding to (2.9), that is, $\phi_i(z, m, s) = \sum_h \psi_i(K + h + 1, m, s)\, z^h$. However, this approach immediately fails, since the stationary probability distribution (3.2) calls for a matrix generalization. The main contribution of this work is to provide this generalization by the introduction of the following matrix generating function,

$$\phi(Z, m, s) = \sum_{h=0}^{\infty} \psi(K + h + 1, m, s)\, Z^h \ , \tag{3.10}$$

where Z is any matrix with eigenvalues contained in the open unit disk of the complex plane.

Remark 3.1. Since the absolute value of the Laplace transform $\psi_i(n, m, s)$ is less than or equal to one, the assumption on the eigenvalues of Z implies that the matrix generating function $\phi(Z, m, s)$ is well defined.

Let us rewrite expression (3.8) for $n > K$ in the alternative form,

$$\psi(n, m, s) = \psi(n-1, m, s)\, T_M(s) + \psi(n, m+1, s)\, T_\Lambda(s) \tag{3.11}$$

with $T_A(s) = A\,(H(s) - \Gamma_1)^{-1}$, $A \in \{\Lambda, M\}$. Multiplying expression (3.11) on the right by Z^h, for $n = K + h + 1$, and then summing over $h \geq 0$ and using that $T\, Z^h\, T^{-1} = (T\, Z\, T^{-1})^h$, we get a recursive equation for $\phi(Z, m, s)$,

$$\begin{aligned} \phi(Z, m, s) &= \psi(K, m, s)\, T_M(s) + \phi(T_M(s)\, Z\, T_M^{-1}(s), m, s)\, T_M(s)\, Z \\ &\quad + \phi(T_\Lambda(s)\, Z\, T_\Lambda^{-1}(s), m+1, s)\, T_\Lambda(s) \ . \end{aligned} \tag{3.12}$$

The main difference between (2.10) and (3.12) is that in the latter we loose the commutative property of the product and the functions ϕ need to be evaluated for different values of their arguments. The boundary condition is obtained from (3.9),

$$\phi(Z, K, s) = e\, T^{K+1}(s) \left(\sum_{h=0}^{\infty} T^h(s)\, Z^h \right) = e\, T^{K+1}(s)\, S(Z, I, T(s)) \tag{3.13}$$

with I being the identity matrix and where we employed the definition

$$S(Z, A, B) := \sum_{h=0}^{\infty} B^h\, A\, Z^h \ . \tag{3.14}$$

The matrix $S(Z, A, B)$ is well defined for any matrix Z, A, B with Z and B having all eigenvalues inside the closed and open disks, respectively (so that the series converges). Note that $T(s)$ in (3.13) has all eigenvalues inside the open unit disk. The matrix $S(Z, A, B)$ can be computed as the solution of a matrix equation as shown in the following lemma. The proof is deferred to the appendix.

LEMMA 3.2. *Let Z, A, and B be three matrices with Z and B having all eigenvalues in the closed and open disks, respectively, then the matrix function $S = S(Z, A, B)$ is the unique solution of the following matrix system,*

$$S - B\,S\,Z = A\ . \tag{3.15}$$

The next proposition shows that, in order to compute the Laplace transform of the stationary sojourn time S^* in terms of a finite number of addends, only the value of $\phi(R, 0, s)$ is needed.

PROPOSITION 3.3. *The Laplace transform of S^* can be computed in the form*

$$\psi(s) = \sum_{n=0}^{K-1} \psi(n+1, 0, s)\,\pi_n + \phi(R, 0, s)\,\pi_K\ . \tag{3.16}$$

Proof. The result follows from (3.6) by splitting the sum into a finite part, $n < K$, and an infinite part. For the latter part, we express π_{K+h} as in (3.2) for $h \geq 0$, and apply definition (3.10). □

The computation of $\phi(R, 0, s)$ requires some additional machinery with respect to the one developed in section 2 for the scalar case. Before giving the statement of the main result we need the following technical lemma, the proof of which is deferred to the appendix. The lemma states that the infinite sum of matrices appearing at the left-hand side of (3.17) can be recognized as a matrix function S, which can be computed from the matrix system (3.15).

LEMMA 3.4. *Let Z, A, and B be three matrices with Z and B having all eigenvalues in the closed and open disks, respectively, and let T_1 and T_2 be invertible matrices with T_1 having all eigenvalues in the open disk, then the following relation holds,*

$$\sum_{h=0}^{\infty} S(T_2\,T_1^h\,Z\,T_1^{-h}\,T_2^{-1}, A, B)\,T_2\,T_1^h\,Z^h = S(Z, A\,T_2\,S(Z, I, T_1), B)\,. \tag{3.17}$$

The following result allows us to compute the value of $\phi(R, m, s)$ in finitely many steps.

THEOREM 3.5. *The values of $\phi(Z, m, s)$ for $0 \leq m \leq K$ can be computed by the following equation*

$$\begin{aligned}\phi(Z, m, s) = &\sum_{k=m}^{K-1} \psi(K, k, s)\,T_M(s)\,U_M(Z, k-m, s) \\ &+ \psi(K, K+1, s)\,T(s)\,U(Z, K-m, s)\ ,\end{aligned} \tag{3.18}$$

where the matrices $U_M(Z, k, s)$ and $U(Z, k, s)$ are defined as

$$\begin{aligned}U_M(Z, k, s) &= S(Z, (T_\Lambda(s)\,S(Z, I, T_M(s)))^k, T_M(s))\ , \\ U(Z, k, s) &= S(Z, (T_\Lambda(s)\,S(Z, I, T_M(s)))^k, T(s))\ .\end{aligned}$$

Proof. Using (3.13) and (3.9), it follows that (3.18) holds for $m = K$, where it is assumed that the value of the sum is zero. We prove by induction that it also holds for all $m < K$. We first derive a recursive equation satisfied by $\phi(\cdot, m, s)$ in terms of $\phi(\cdot, m+1, s)$.

By substituting $T_M(s)\,Z\,T_M^{-1}(s)$ for Z in (3.12) we get an expression for $\phi(T_M(s)\,ZT_M^{-1}(s), m, s)$, and subsequently substituting this expression in the right-

hand side of (3.12), yields

$$\begin{aligned}\phi(Z,m,s) &= \psi(K,m,s)\,T_M(s) + \psi(K,m,s)\,T_M^2(s)\,Z \\ &\quad + \phi(T_M^2(s)\,Z\,T_M^{-2}(s),m,s)\,T_M^2(s)\,Z^2 \\ &\quad + \phi(T_\Lambda(s)\,T_M(s)\,Z\,T_M^{-1}(s)\,T_\Lambda^{-1}(s),m+1,s)\,T_\Lambda(s)\,T_M(s)\,Z \\ &\quad + \phi(T_\Lambda(s)\,Z\,T_\Lambda^{-1}(s),m+1,s)\,T_\Lambda(s) \qquad (3.19)\end{aligned}$$

and iterating this equation leads to

$$\begin{aligned}\phi(Z,m,s) &= \psi(K,m,s)\,T_M(s)\,\left(\sum_{h=0}^{\infty} T_M^h(s)\,Z^h\right) \\ &\quad + \sum_{h=0}^{\infty} \phi(T_\Lambda(s)\,T_M^h(s)\,Z\,T_M^{-h}(s)\,T_\Lambda^{-1}(s),m+1,s)\,T_\Lambda(s)\,T_M^h(s)\,Z^h\ ,\end{aligned}$$

which can be rewritten as

(3.20)

$$\begin{aligned}\phi(Z,m,s) &= \psi(K,m,s)\,T_M(s)\,S(Z,I,T_M) \\ &\quad + \sum_{h=0}^{\infty} \phi(T_\Lambda(s)\,T_M^h(s)\,Z\,T_M^{-h}(s)\,T_\Lambda^{-1}(s),m+1,s)\,T_\Lambda(s)\,T_M^h(s)\,Z^h\ .\end{aligned}$$

The recursive equation (3.20) is valid for $m = 0, 1, \ldots, K-1$.

We conjecture that for all $m = 0, 1, \ldots, K$, the generating function $\phi(Z,m,s)$ has the form

$$\begin{aligned}\phi(Z,m,s) &= \sum_{k=m}^{K-1} \psi(K,k,s)\,T_M(s)\,S(Z,Y^{k-m}(s),T_M(s)) \\ &\quad + \psi(K,K+1,s)\,T(s)\,S(Z,Y^{K-m}(s),T(s))\ ,\end{aligned} \qquad (3.21)$$

so that (3.18) follows by showing that the right expression for $Y(s)$ is given by

$$Y(s) = T_\Lambda(s)\,S(Z,I,T_M(s))\ . \qquad (3.22)$$

This conjecture will be proved by induction. We have already shown that it holds for $m = K$. Now assume that it is valid for $m+1$. To establish (3.21) for m, it suffices to prove, by virtue of (3.20), that

$$\begin{aligned}&\sum_{h=0}^{\infty} \phi(T_\Lambda(s)\,T_M^h(s)\,Z\,T_M^{-h}(s)\,T_\Lambda^{-1}(s),m+1,s) T_\Lambda(s)\,T_M^h(s)\,Z^h \\ &= \sum_{k=m+1}^{K-1} \psi(K,k,s)\,T_M(s)\,S(Z,Y^{k-m}(s),T_M(s)) \\ &\quad + \psi(K,K+1,s)\,T(s)\,S(Z,Y^{K-m}(s),T(s))\ .\end{aligned} \qquad (3.23)$$

It follows from Lemma 3.4 that

$$\begin{aligned}&\psi(K,K+1)\,T\sum_{h=0}^{\infty} S(T_\Lambda\,T_M^h\,Z\,T_M^{-h}\,T_\Lambda^{-1},Y^{K-m-1},T)\,T_\Lambda\,T_M^h\,Z^h \\ &\qquad = \psi(K,K+1)\,TS(Z,\,Y^{K-m-1}\,T_\Lambda\,S(Z,I,T_M),T)\ ,\end{aligned} \qquad (3.24)$$

where we suppressed the dependence on s. Application of Lemma 3.4 is justified, since it is readily verified that the matrices in the above infinite sum satisfy the conditions

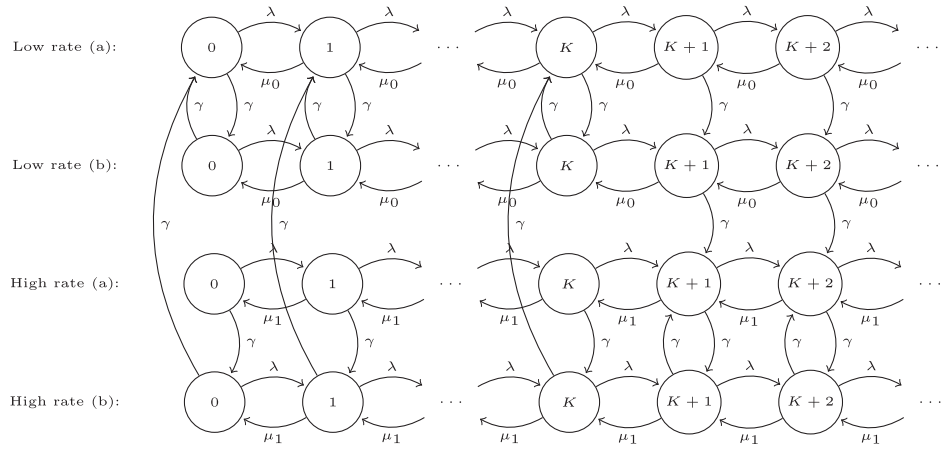

FIG. 4. *Transition diagram for Erlang-2 inspection times.*

mentioned in this lemma. Accordingly, for $k = m+1, \ldots, K-1$, and again suppressing the dependence on s,

$$\psi(K,k)\, T_M \sum_{h=0}^{\infty} S(T_\Lambda\, T_M^h\, Z\, T_M^{-h}\, T_\Lambda^{-1}, Y^{k-m-1}, T_M)\, T_\Lambda\, T_M^h\, Z^h$$

$$= \psi(K,k)\, T_M\, S(Z, Y^{k-m-1}\, T_\Lambda\, S(Z, I, T_M), T_M)\ . \tag{3.25}$$

Combining (3.24) and (3.25) we conclude, by virtue of the induction hypothesis, that (3.23) holds whenever $Y(s)$ satisfies (3.22), which completes the proof. □

Remark 3.6. Also in this case, as was already mentioned in Remark 2.4, the Laplace transform of the sojourn time is rational. This admits application of classical inversion techniques, yielding an explicit expression for the sojourn time distribution. In section 3.2 we give an example of how to compute the density function of the sojourn time for a system with $K = 2$.

3.1. Erlang inspection times. In section 3 we assumed exponential interinspection times. In principle this can be extended to the case of phase-type distributed interinspection times [1], paying a cost in terms of model complexity. Indeed, in this case one should keep track, not only of the value of the service rate, but also of the phase of the inspection clock. This translates into more complicated matrix expressions, but the basic logic of the computation of the sojourn time distribution remains the same. In fact, this is exactly the power of the proposed matrix generating function technique. For the sake of clarity and conciseness we are not going to treat here this extension in detail, but give a quick view of how it can be handled.

We assume that the inspection times are Erlang(2,γ) distributed. To keep trace of this we consider four states in the description of the system, $\{00, 01, 10, 11\}$, where the first number specifies the speed of the system and the second the phase of the inspection clock (see the transition diagram in Figure 4).

The column vector $\pi_n = (\pi_{00n}, \ldots, \pi_{11n})^\top$ satisfies (3.1) with the following matrices

$$M = \begin{pmatrix} \mu_0 & 0 \\ 0 & \mu_1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\ , \quad \Lambda = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\ ,$$

and $H_1 = \Lambda + \Gamma_2$, $H_2 = M + \Lambda + \Gamma_2$, and $H_3 = M + \Lambda + \Gamma_3$, where

$$\Gamma_2 = \begin{pmatrix} \gamma & -\gamma & 0 & -\gamma \\ -\gamma & \gamma & 0 & 0 \\ 0 & 0 & \gamma & 0 \\ 0 & 0 & -\gamma & \gamma \end{pmatrix}, \quad \Gamma_3 = \begin{pmatrix} \gamma & 0 & 0 & 0 \\ -\gamma & \gamma & 0 & 0 \\ 0 & -\gamma & \gamma & -\gamma \\ 0 & 0 & -\gamma & \gamma \end{pmatrix}.$$

The conditional sojourn times satisfy the following equation (see the corresponding formula (3.7)),

$$S_{ij}(n,m) = \frac{X}{\lambda + \mu_i + \gamma} + \begin{cases} S_{i,j}(n-1,m) & \text{w.p. } \mu_i/(\lambda+\mu_i+\gamma), \\ S_{i,j}(n,m+1) & \text{w.p. } \lambda/(\lambda+\mu_i+\gamma), \\ S_{h(i,j)}(n,m) & \text{w.p. } \gamma/(\lambda+\mu_i+\gamma) \end{cases} \tag{3.26}$$

with $h(i,j) = h(i,j;n,m) = ((1-j)\cdot i + j\cdot 1\{n+m > K\}, (1-j))$. It follows that the row vector $(\psi_{00}(n,m,s), \psi_{01}(n,m,s), \psi_{10}(n,m,s), \psi_{11}(n,m,s))$ satisfies (3.8) with the matrices $H(s) = (s+\gamma)\, I + \Lambda + M$ and

$$\Gamma_0 = \begin{pmatrix} 0 & \gamma & 0 & \gamma \\ \gamma & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \gamma & 0 \end{pmatrix}, \quad \Gamma_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \gamma & 0 & 0 & 0 \\ 0 & \gamma & 0 & \gamma \\ 0 & 0 & \gamma & 0 \end{pmatrix}.$$

Since the matrix equations for the Erlang inspection times are similar to the exponential inspection times, all the subsequent matrix analysis in Proposition 3.3 and Theorem 3.5 are still valid. The value of the matrix R is now given by

$$R = \begin{pmatrix} R_{11} & 0 & 0 & 0 \\ R_{21} & R_{11} & 0 & 0 \\ R_{31} & R_{32} & R_{33} & R_{34} \\ R_{41} & R_{42} & R_{34} & R_{33} \end{pmatrix}$$

with

$$R_{11} = \frac{\gamma + \lambda + \mu_0 - \sqrt{-4\lambda\mu_0 + (\gamma+\lambda+\mu_0)^2}}{2\mu_0}, \quad R_{21} = \frac{\gamma R_{11}}{\gamma + \lambda + \mu_0 - 2\mu_0 R_{11}},$$

$$R_{33} = \frac{\mu_1(2\gamma + 3\lambda + \mu_1) - \sqrt{\mu_1^2\left(4\gamma^2 + (\lambda-\mu_1)^2 + 4\gamma(\lambda+\mu_1)\right)}}{4\mu_1^2},$$

$$R_{34} = \frac{\lambda}{\mu_1} - R_{33}, \quad R_{31} = \frac{\gamma(R_{41}+R_{21}) + \mu_1(R_{32}R_{21} + R_{41}R_{34})}{\gamma + \lambda - \mu_1(-1 + R_{11} + R_{33})},$$

$$R_{32} = \frac{\gamma R_{11}(-\gamma - \lambda + \mu_1(-1 + R_{11} + R_{33}))}{-(\gamma + \lambda - \mu_1(-1 + R_{11} + R_{33}))^2 + (\gamma + \mu_1 R_{34})^2},$$

$$R_{41} = \frac{\gamma R_{21}(\gamma + \mu_1 R_{34})\left(\lambda^2 - 2\lambda\mu_1(-1+R_{33}) - \mu_1^2\left(R_{11}^2 - (-1+R_{33})^2 + R_{34}^2\right)\right)}{(2\gamma + \lambda - \mu_1(-1+R_{11}+R_{33}-R_{34}))^2(\lambda - \mu_1(-1+R_{11}+R_{33}+R_{34}))^2}$$
$$\times \frac{\gamma R_{21}(\gamma+\mu_1 R_{34})\left(2\gamma(\lambda - \mu_1(-1+R_{33}+R_{34}))\right)}{(2\gamma + \lambda - \mu_1(-1+R_{11}+R_{33}-R_{34}))^2(\lambda - \mu_1(-1+R_{11}+R_{33}+R_{34}))^2},$$

$$R_{42} = \frac{\gamma R_{11}(\gamma + \mu_1 R_{34})}{(2\gamma + \lambda - \mu_1(-1+R_{11}+R_{33}-R_{34}))(\lambda - \mu_1(-1+R_{11}+R_{33}+R_{34}))}.$$

3.2. Analytical example. In this section we briefly show that by using Theorem 3.5, we can get explicit expressions for the density function of the sojourn time in the system with inspection times, as highlighted in Remark 3.6.

The computations are simple, but tedious as they require extensive use of matrix calculus, and usually it is easy when assisted by symbolic computational software as we do for this example.

To make the computations easier, we wisely select the values of the parameters of the system such that all coefficients turn out to be rational.

The parameters of the queue are

$$\mu_0 = 1, \quad \mu_1 = 3/2, \quad \lambda = 9/8, \quad \gamma = 1/8 \; .$$

For the moment we do not fix the threshold; later we consider explicitly the case $K = 2$. The above choice of the parameters gives $R_{00} = 3/4$ in (3.4). The matrix R and the matrix function $T(s) = M\,(H(s) - \Gamma_1 - \Lambda)^{-1}$, are given by

$$R = \begin{pmatrix} \frac{3}{4} & 0 \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix}, \quad T(s) = \begin{pmatrix} \frac{8}{9+8s} & 0 \\ \frac{3}{(3+2s)(9+8s)} & \frac{3}{3+2s} \end{pmatrix},$$

and the matrix functions $T_\Lambda(s) = \Lambda\,(H(s) - \Gamma_1)^{-1}$ and $T_M(s) = M\,(H(s) - \Gamma_1)^{-1}$ are equal to

$$T_\Lambda(s) = \begin{pmatrix} \frac{9}{2(9+4s)} & 0 \\ \frac{9}{2(9+4s)(21+8s)} & \frac{9}{21+8s} \end{pmatrix}, \quad T_M(s) = \begin{pmatrix} \frac{4}{9+4s} & 0 \\ \frac{6}{(9+4s)(21+8s)} & \frac{12}{21+8s} \end{pmatrix} .$$

Solving the matrix system (3.15) we get the following expression for $S(R, I, T_M(s))$,

$$S(R, I, T_M(s)) = \begin{pmatrix} \frac{9+4s}{2(3+2s)} & 0 \\ \frac{3(3+s)}{2(3+2s)^2} & \frac{21+8s}{4(3+2s)} \end{pmatrix}$$

that allows the computation of the values of $U(R, k, s)$ and $U_M(R, k, s)$ for any $k \geq 0$. As an example we show such matrix functions for $k = 2$,

$$U(R, 2, s) = \begin{pmatrix} \frac{81(9+8s)}{16(3+2s)^2(3+8s)} & 0 \\ \frac{81(69+88s)}{16(3+2s)^2(3+8s)^2} & \frac{81}{4(3+2s)(3+8s)} \end{pmatrix},$$

$$U_M(R, 2, s) = \begin{pmatrix} \frac{81(9+4s)}{32(3+2s)^3} & 0 \\ \frac{81(30+11s)}{32(3+2s)^4} & \frac{81(21+8s)}{64(3+2s)^3} \end{pmatrix} .$$

Remark 3.7. The expressions for $S(R, I, T_M(s)))$, $U(R, k, s)$, and $U_M(R, k, s)$ do not depend on K, so they can be used for any value of the threshold. The values of $\psi(s)$, $\psi(n, 0, s)$, π_n, and $\phi(R, 0, s)$ in (3.16) do depend on K via the respective formulas (3.16), (3.8), (3.5), and (3.18).

From here on we fix $K = 2$. We have $\pi_K = (3807/60644, 1701/30322)^\top$ and after recursively computing $\psi(k, 0, s)$, for $k = 1, 2$, we finally get $\psi(s)$,

$$\begin{aligned} \psi(s) = &-\frac{308367}{379025(3+2s)^4} - \frac{13923657}{9475625(3+2s)^3} - \frac{44764461}{47378125(3+2s)^2} \\ &- \frac{130808703}{236890625(3+2s)} - \frac{14013}{15161(9+4s)} + \frac{1587762}{9475625(11+4s)^3} \\ &- \frac{4755267}{24636625(11+4s)^2} - \frac{28797784929}{40034515625(11+4s)} + \frac{81216}{15161(3+8s)^2} \\ &+ \frac{18144}{15161(3+8s)} + \frac{102060}{15161(9+8s)^2} + \frac{55081053}{20497672(9+8s)} - \frac{24064452}{9475625(17+8s)^3} \\ &- \frac{199526994}{47378125(17+8s)^2} + \frac{2950774277}{1895125000(17+8s)} + \frac{90111}{60644(21+8s)} \; , \end{aligned}$$

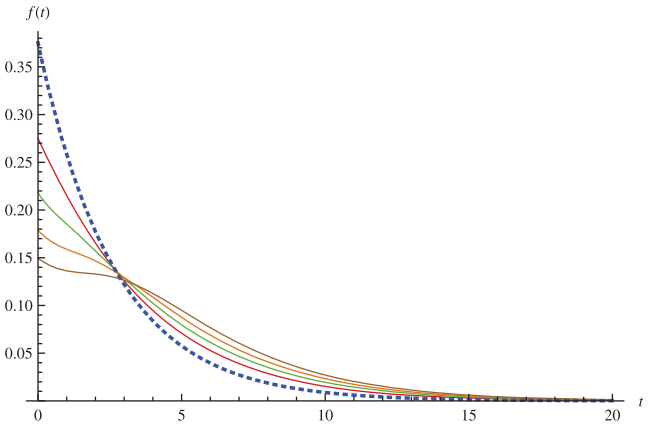

FIG. 5. *Density function $f(t)$ of the sojourn time S^* for $\mu_0 = 1$, $\mu_1 = 3/2$, $\lambda = 9/8$, $\gamma = 1/8$; $K = 0$ is the red line, $K = 1$ is the green line, $K = 2$ is the orange line, and $K = 3$ is the brown line. The dashed blue line shows the case of a classical $M/M/1$ queue with service rate μ_1.*

the inverse transform of which results into the following density function

$$\begin{aligned} f(t) = &-\frac{90111e^{-21t/8}}{485152} - \frac{14013e^{-9t/4}}{60644} + \frac{27e^{-3t/8}(84+47t)}{15161} \\ &+ \frac{729e^{-9t/8}(75557+23660t)}{163981376} \\ &+ 243e^{-11t/4}\left(-1896150448 - 127198500t + 13803075t^2\right)/2562209000000 \\ &- e^{-17t/8}\left(-11803097108 + 3990539880t + 150402825t^2\right)/60644000000 \\ &- 3e^{-3t/2}\left(697646416 + 596859480t + 232060950t^2 + 21414375t^3\right)/7580500000. \end{aligned}$$

Figure 5 plots the density functions of the sojourn time for $K = 0, 1, 2, 3$ using their exact expressions, instead of using the numeric inverse transform as is done later on in section 4.

3.3. First moment calculation. As in the previous section, we define $\nu = \mathbb{E}[S]$ and $\nu_{n,m} = \mathbb{E}[S(n,m)]$. By taking derivatives in (3.8) and then computing the limit for $s \to 0$ we get

$$\nu_{n,m}\left(H(0) - \Gamma_{1\{n+m>K\}}\right) = \nu_{n-1,m}\, M + \nu_{n,m+1}\, \Lambda + e\ , \tag{3.27}$$

where we used that $H'(0)$ is the identity matrix. The vector e is the all-one vector. From (3.9) and after taking derivatives, we obtain

$$\nu_{n,m} = e \sum_{k=1}^{n} (T(0))^k\, M^{-1}\, (T(0))^{n-k+1} \quad \text{as } n > 0 \text{ and } m \geq K\ , \tag{3.28}$$

with $T(0) = M\,(H(0) - \Gamma_1 - \Lambda)^{-1}$, $(T^{-1})'(0) = M^{-1}$, and $T'(0) = -T(0)M^{-1}T(0)$. Here we used that the derivative of a matrix A^{-n} is given by

$$(A^{-n})' = \sum_{k=1}^{n} A^{-k}\, A' A^{k-n-1}\ .$$

By Proposition 3.3 we can conclude that

$$\nu = \sum_{n=0}^{K-1} \nu_{n+1,0}\, \pi_n - \phi'(Z,0,0+)\, \pi_K \ . \tag{3.29}$$

From equation (3.18) we can compute

$$\begin{aligned} -\phi'(Z,0,0+) = & \sum_{k=0}^{K-1} \nu_{K,k}\, T_M(0)\, U_M(Z,k,0) + \nu_{K,K+1}\, T(0)\, U(Z,K,0) \\ & - e \sum_{k=0}^{K-1} T_M(0)\, U'_M(Z,k,0) - e\, T(0)\, U'(Z,K,0) \\ & - e \sum_{k=0}^{K-1} T'_M(0)\, U_M(Z,k,0) - e\, T'(0)\, U(Z,K,0) \ , \end{aligned} \tag{3.30}$$

with $T_M(0) = M\,(H(0) - \Gamma_1)^{-1}$ and $T'_M(0) = T_M(0)\, M^{-1}\ T_M(0)$. The values $U'_M(Z,k,0)$ and $U'(Z,k,0)$ appearing in (3.18) can be computed by solving the following linear systems, see Lemma A.1 in the appendix,

$$\begin{aligned} U'_M(Z,k,0) - T_M(0)\, U'_M(Z,k,0)\, Z - T'_M(0)\, U_M(Z,k,0)\, Z &= A'(Z,k,0) \\ U'(Z,k,0) - T(0)\, U'(Z,k,0)\, Z - T'(0)\, U(Z,k,0)\, Z &= A'(Z,k,0) \end{aligned}$$

with $A(Z,k,s) = (T_\Lambda(s)\, S(Z,I,T_M(s)))^k$.

4. Numerical experiments. In this section we show some numerical examples, where we compute the stationary sojourn time distribution for a system with slow rate $\mu_0 = 1$ and high rate $\mu_1 = 3/2$.

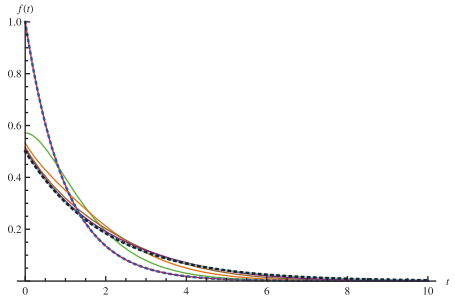

FIG. 6. $\lambda = 1/2$

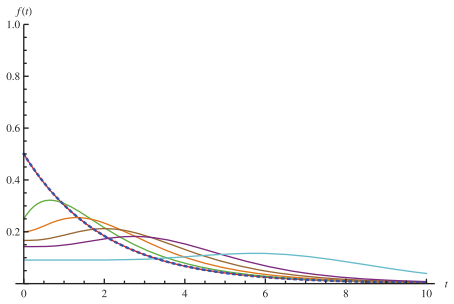

FIG. 7. $\lambda = 1$

Density function $f(t)$ of the sojourn time S^* for $\mu_0 = 1$, $\mu_1 = 3/2$; $K = 0$ is the red line, $K = 1$ is the green line, $K = 2$ is the orange line, $K = 3$ is the brown line, $K = 4$ is the purple line, and $K = 5$ is the cyan line. For $\lambda < \mu_0$, the dashed black line shows the case of a classical $M/M/1$ queue with service rate μ_0 (equivalent to setting $K = \infty$). The dashed blue line shows the case of a classical $M/M/1$ queue with service rate μ_1.

In Figures 6 and 7, it is shown how the sojourn time distribution depends on the threshold K for the case of immediate switching times. In the first example, $\lambda < \mu_0 < \mu_1$, which implies that the system is stable for both service rates. Therefore,

when $K \to \infty$, one can appreciate that the sojourn time distribution approaches the one of an $M/M/1$ system with fixed service rate μ_0 (shown as dashed black line in Figure 6). In the second example, we have $\lambda \in [\mu_0, \mu_1)$. In particular, we have chosen $\lambda = \mu_0 = 1$, implying that the system approaches instability as $K \to \infty$.

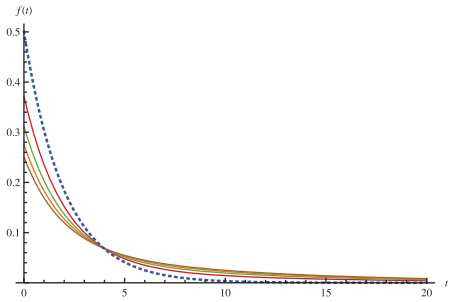

FIG. 8. $\gamma = 1/100$

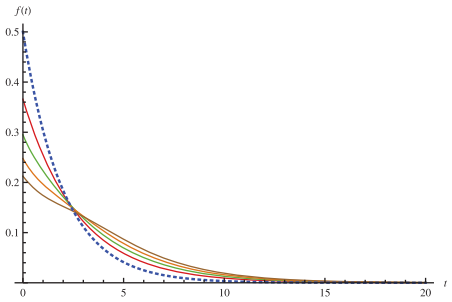

FIG. 9. $\gamma = 1/10$

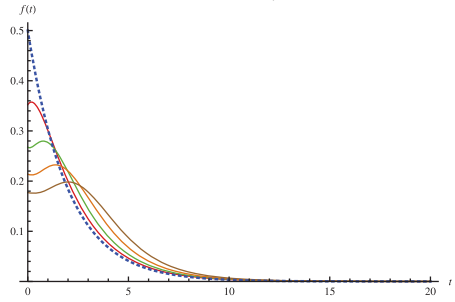

FIG. 10. $\gamma = 1$

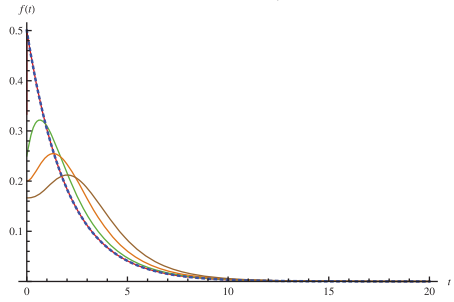

FIG. 11. $\gamma = 1000$

Density function $f(t)$ of the sojourn time S^* for $\mu_0 = 1$, $\mu_1 = 3/2$, $\lambda = 1$; $K = 0$ is the red line, $K = 1$ is the green line, $K = 2$ is the orange line, and $K = 3$ is the brown line. The dashed blue line shows the case of a classical $M/M/1$ queue with service rate μ_1.

Figures 8–11 show the sojourn time distribution for the case of exponential distributed inspection times. These figures refer to the case when $\lambda \in [\mu_0, \mu_1)$, and again, one can notice that as $K \to \infty$, the system becomes unstable. It is worth noticing that, when $K = 0$, the curve does not coincide with the $M/M/1$ with constant service rate μ_1 (shown as dashed blue line), since in the system with exponential switching, when inspection finds the system empty, the server switches to the slow rate and does not switch back till another inspection occurs. When $\gamma = 1000$, the system switches almost immediately and therefore the sojourn time distribution is very close to the one of the pure $M/M/1$ system.

In Figures 12 and 13, we plot again the results for $\lambda = 1$, $\mu_0 = 1$, and $\mu_1 = 3/2$, but compare different values of γ's. One can see that for γ approaching λ, the system behaves very closely to a system with immediate switching (shown as dashed black line). Indeed, for values of $\gamma > 1$, one cannot distinguish the curve from the limiting one. This suggests that, checking the state of the system at a rate comparable to the arrival rate can be considered from the point of view of the sojourn time as an immediate switching. This could be used in the design phase of the system, when balancing between costs (by reducing service rate) and performance (by increasing the service and inspection rate).

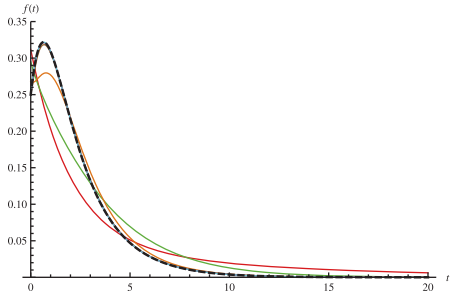

FIG. 12. $K = 1$

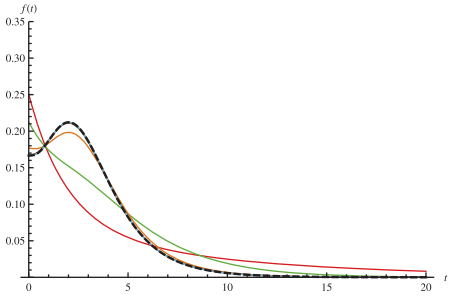

FIG. 13. $K = 3$

Density function $f(t)$ of the sojourn time S^* for $\mu_0 = 1$, $\mu_1 = 3/2$, $\lambda = 1$; $\gamma = 1/100$ is the red line, $\gamma = 1/10$ is the green line, $\gamma = 1$ is the orange line, $\gamma = 10$ is the brown line, $\gamma = 100$ is the purple line, $\gamma = 1000$ is the cyan line. The dashed black line shows the case of continuous inspection (equivalent to setting $\gamma = \infty$).

5. Conclusions. In this paper we studied the sojourn time distribution in an exponential single-server queueing system. Service is in order of arrival, and it is provided at low or high rate, which can be adapted at exponential inspection times, depending on the number of customers in the system. To determine the Laplace transform of the stationary sojourn time distribution, we proposed a new methodological tool, that is matrix generating functions. We used this tool to compute the Laplace transform of the sojourn time distribution in the system with inspection times. Its expression is obtained recursively and shows a rational form that allows an immediate inverse transformation. Numerical computations have shown, as expected, that if the inspection rate is large, the sojourn time of the system with inspections converges to the one of the system with immediate switching.

We believe that the power of the matrix generating functions lies in its flexibility to analyze generalizations to phase-type services and inspection times. An interesting and promising direction for future research is to explore the applicability of this tool to analyze the more general class of quasi-birth-and-death processes [13].

Appendix A. Technical proofs.

Proof of Lemma 2.1. We can rewrite the expression (2.5) in the following form

$$\psi(n,m,s) = a_s(n+m)\,\psi(n-1,m,s) + b_s(n+m)\,\psi(n,m+1,s)\ . \tag{A.1}$$

With $m = K-1$, (2.7) becomes

$$\begin{aligned}\psi(n,K-1,s) &= B_s(n+K-1,1)\,\psi(n,K,s)\\ &\quad + \sum_{k=K-1}^{K-1} a_s(n+k)B_s(n+K-1,k-K+1)\,\psi(n-1,k,s)\\ &= b_s(n+K-1)\,\psi(n,K,s) + a_s(n+K-1)\,\psi(n-1,K-1,s)\end{aligned}$$

and therefore it holds true. Now assume that (2.7) is valid for $m+1$. Then by (A.1),

$$
\begin{aligned}
\psi(n,m,s) &= a_s(n+m)\,\psi(n-1,m,s) \\
&\quad + b_s(n+m)\,B_s(n+m+1,K-1-m)\,\psi(n,K,s) \\
&\quad + b_s(n+m)\sum_{k=m+1}^{K-1} a_s(n+k)\,B_s(n+m+1,k-m-1)\,\psi(n-1,k,s) \\
&= a_s(n+m)\,B_s(n+m,0)\,\psi(n-1,m,s) + B_s(n+m,K-m)\,\psi(n,K,s) \\
&\quad + \sum_{k=m+1}^{K-1} a_s(n+k)\,B_s(n+m,k-m)\,\psi(n-1,k,s) \\
&= B_s(n+m,K-m)\,\psi(n,K,s) \\
&\quad + \sum_{k=m}^{K-1} a_s(n+k)\,B_s(n+m,k-m)\,\psi(n-1,k,s)\ ,
\end{aligned}
$$

where we have used the fact that the definition of $B_s(k,h)$ implies that

$$b_s(k)\,B_s(k+1,h) = B_s(k,h+1)\ . \qquad \square$$

Proof of Lemma 3.2. By substituting in (3.15) the expression for S given in (3.14) we get

$$B\,S\,Z = \sum_{h=0}^{\infty} B^{h+1}\,A\,Z^{h+1} = \sum_{h=0}^{\infty} B^h\,A\,Z^h\ - A = S - A$$

which implies that the matrix S is a solution of the matrix equation.

By assuming that S and S' are two solutions of this matrix equation, we would have that $Y = S - S'$ is the solution of the following system

$$Y = Z\,Y\,B.$$

Iterating the last equation we get that

$$Y = Z^n\ Y\,B^n, \quad n \geq 0\ .$$

This term converges to 0 as $n \to \infty$ by the assumptions on the eigenvalues of the matrices Z and B. It follows that $Y = 0$ and hence S is unique. $\square$

Proof of Lemma 3.4. The result follows from the following algebraic manipulations

$$
\begin{aligned}
&\sum_{h=0}^{\infty} S(T_2\,T_1^h\,Z\,T_1^{-h}\,T_2^{-1},A,B)\,T_2\,T_1^h\,Z^h \\
&\quad = \sum_{h=0}^{\infty}\sum_{k=0}^{\infty} B^k\,A\,(T_2\,T_1^h\,Z\,T_1^{-h}\,T_2^{-1})^k\,T_2\,T_1^h\,Z^h \\
&\quad = \sum_{h=0}^{\infty}\sum_{k=0}^{\infty} B^k\,A\,T_2\,T_1^h\,Z^k\,T_1^{-h}\,T_2^{-1}\,T_2\,T_1^h\,Z^h \\
&\quad = \sum_{h=0}^{\infty}\sum_{k=0}^{\infty} B^k\,A\,T_2\,T_1^h\,Z^{k+h} = \sum_{k=0}^{\infty} B^k\,A\,T_2\left(\sum_{h=0}^{\infty} T_1^h\,Z^h\right)Z^k \\
&\quad = \sum_{k=0}^{\infty} B^k\,A\,T_2\,S(Z,I,T_1)\,Z^k = S(Z,A\,T_2\,S(Z,I,T_1),B)\,. \qquad \square
\end{aligned}
$$

LEMMA A.1. *Let $S(s) = S(Z, A(s), B(s))$, then its derivative in s can be computed as the solution of the following linear system*

$$S'(s) - B(s)\,S'(s)\,Z - B'(s)\,S(s)\,Z = A'(s)\,. \tag{A.2}$$

Proof. By (3.15) we have that

$$S(Z, A(s+h), B(s+h)) - B(s+h)\,S(Z, A(s+h), B(s+h))\,Z = A(s+h)\,, \tag{A.3}$$

$$S(Z, A(s), B(s)) - B(s)\,S(Z, A(s), B(s))\,Z = A(s)\ . \tag{A.4}$$

Subtracting the expressions above, adding and removing $B(s+h)\,S(Z, A(s), B(s))\,Z$ we have

$$\Delta S(s) - B(s+h)\,\Delta S(s)\,Z - \Delta B(s)\,S(Z, A(s), B(s))\,Z = \Delta A(s)$$

with $\Delta S(s) = S(Z, A(s+h), B(s+h)) - S(Z, A(s), B(s))$ and similar notations for $\Delta A(s)$ and $\Delta B(s)$. Dividing by h and letting $h \to 0$ the result follows. ∎

REFERENCES

[1] S. ASMUSSEN, *Applied Probability and Queues*, Springer, New York, 2003.

[2] R. BEKKER, S.C. BORST, O.J. BOXMA, AND O. KELLA, *Queues with workload-dependent arrival and service rates*, Queueing Syst., 46 (2004), pp. 537–556.

[3] R. BEKKER AND O.J. BOXMA, *An $M/G/1$ queue with adaptable service speed*, Stoch. Models, 23 (2007), pp. 373–396.

[4] R. BEKKER, O.J. BOXMA, AND J.A.C. RESING, *Queues with adaptable service speed*, Stat. Neerl., 62 (2008), pp. 441–457.

[5] O.J. BOXMA, B.H.B. JONSSON, J.A.C. RESING, AND V. SHNEER, *An alternating risk reserve process - Part* II, Markov Process. Related Fields, 16 (2010), pp. 425–446.

[6] J.W. COHEN, *On the optimal switching level for an $M/G/1$ queueing system*, Stochastic Process. Appl., 4 (1976), pp. 297–316.

[7] J.W. COHEN, *The Single Server Queue*, North-Holland, Amsterdam, 1982.

[8] G.I. FALIN, *On the waiting-time process in a single-line queue with repeated calls*, J. Appl. Probab., 23 (1986), pp. 185–192.

[9] G.I. FALIN, *A survey of retrial queues*, Queueing Syst., 7 (1990), pp. 127–167.

[10] D.P. GAVER AND R.G. MILLER, *Limiting distributions for some storage problems*, in Studies in Applied Probability and Management Science, Stanford University Press, Stanford, CA, 1962, pp. 110–126.

[11] J.M. HARRISON AND S.I. RESNICK, *The stationary distribution and first exit probabilities of a storage process with general release rule*, Math. Oper. Res., 1 (1976), pp. 347–358.

[12] J. KEILSON AND L.D. SERVI, *A distributional form of Little's Law*, Oper. Res. Lett., 7 (1988), pp. 223–227.

[13] G. LATOUCHE AND V. RAMASWAMI, *Introduction to Matrix Analytic Methods in Stochastic Modeling*, SIAM, Philadelphia, 1999.

[14] I. MITRANI, *Managing performance and power consumption in a server farm*, Ann. Oper. Res., 202 (2013), pp. 121–134.

[15] M.F. NEUTS, *Matrix-Geometric Solutions in Stochastic Models - An Algorithmic Approach*, Dover, New York, 1994.

[16] R.W. WOLFF, *Poisson arrivals see time averages*, Oper. Res., 30 (1982), pp. 223–231.